Isaac Asimovs robotlove

Mennesker udfører fra tid til anden handlinger, der er til skade for dem selv, andre eller endda hele menneskeheden. Hvordan sikrer vi os mod at udvikle kunstigt intelligente systemer, der vil kunne gøre det samme? Isaac Asimovs robotlove er et bud herpå.

Adfærdsregler for robotter

Kunstigt intelligente systemer er elektriske og/eller mekaniske systemer, der er i stand til at lære nye ting, korrigere sig selv og træffe selvstændige beslutninger. Mange forestiller sig, at kunstig intelligens vil ligne menneskelig intelligens. Men mennesker er ikke altid gode. De kan skade sig selv, andre mennesker og menneskeheden. Hvad vil forhindre robotter, computere og andre systemer med kunstig intelligens i at gøre det samme?

Den amerikansk-russiske biokemiker og science fiction-forfatter Isaac Asimov skrev gerne historier om sympatiske robotter og mente, at skrækscenarier om fremtidens robotteknologi var trættende. I 1942 introducerede Asimov for første gang det, der siden hen er blevet kendt som ”de tre robotlove”. Lovene angiver følgende regler, som enhver kunstigt intelligent robot må overholde:

1. lov: En robot må ikke skade et menneske eller ved uvirksomhed lade et menneske komme til skade.

2. lov: En robot skal adlyde ordrer, som gives til den af mennesker, for så vidt disse ordrer ikke er i konflikt med den første lov.

3. lov: En robot skal beskytte sin egen eksistens, for så vidt dette ikke er i konflikt med den første eller den anden lov.

Senere blev en nulte lov tilføjet:

0. lov: En robot må ikke skade menneskeheden eller ved uvirksomhed lade menneskeheden blive skadet.

Lovene fungerer hierarkisk. Det vil sige, at den nulte lov er den vigtigste, mens den første lov er vigtigere end den anden, ligesom den første og den anden lov er vigtigere end den tredje. I praksis betyder det, at en robot for eksempel ikke kan beordres til at skade eller dræbe et menneske, medmindre det er for at redde menneskeheden (idet den første lov trumfer den anden og den nulte lov trumfer den første), men godt kan beordres til at selvdestruere (da den anden lov trumfer den tredje).

På denne måde fungerer Asimovs love som en slags robotetik, der sikrer, at kunstigt intelligente systemer kan frembringes og fungere uden fare for mennesker og menneskeheden. Formålet med lovene er med andre ord at beskytte mennesker fra at skabe kunstigt intelligente robotter og systemer, der kan være til fare for deres menneskelige omgivelser.

Robotlove parallel til menneskelig adfærd

Kunstigt intelligente systemer er elektriske og/eller mekaniske systemer, der er i stand til at lære nye ting, korrigere sig selv og træffe selvstændige beslutninger. Mange forestiller sig, at kunstig intelligens vil ligne menneskelig intelligens. Men mennesker er ikke altid gode. De kan skade sig selv, andre mennesker og menneskeheden. Hvad vil forhindre robotter, computere og andre systemer med kunstig intelligens i at gøre det samme? Den amerikansk-russiske biokemiker og science fiction-forfatter Isaac Asimov skrev gerne historier om sympatiske robotter og mente, at skrækscenarier om fremtidens robotteknologi var

trættende. I 1942 introducerede Asimov for første gang det, der siden hen er blevet kendt som "de tre robotlove". Lovene angiver følgende regler, som enhver kunstigt intelligent robot må overholde: 1. lov: En robot må ikke skade et menneske eller ved uvirksomhed lade et menneske komme til skade. 2. lov: En robot skal adlyde ordrer, som gives til den af mennesker, for så vidt disse ordrer ikke er i konflikt med den første lov. 3. lov: En robot skal beskytte sin egen eksistens, for så vidt dette ikke er i konflikt med den første eller den anden lov. Senere blev en nulte lov tilføjet: 0. lov: En robot må ikke skade menneskeheden eller ved uvirksomhed lade menneskeheden blive skadet. Lovene fungerer hierarkisk. Det vil sige, at den nulte lov er den vigtigste, mens den første lov er vigtigere end den anden, ligesom den første og den anden lov er vigtigere end den tredje. I praksis betyder det, at en robot for eksempel ikke kan beordres til at skade eller dræbe et menneske, medmindre det er for at redde menneskeheden (idet den første lov trumfer den anden og den nulte lov trumfer den første), men godt kan beordres til at selvdestruere (da den anden lov trumfer den tredje). På denne måde fungerer Asimovs love som en slags robotetik, der sikrer, at kunstigt intelligente systemer kan frembringes og fungere uden fare for mennesker og menneskeheden. Formålet med lovene er med andre ord at beskytte mennesker fra at skabe kunstigt intelligente robotter og systemer, der kan være til fare for deres menneskelige omgivelser.

Samtidig kan man se lovene som en slags parallel til, hvordan mennesker i øvrigt forventes at omgås hinanden. Med undtagelse af ekstreme tilfælde som krig og lignende forventes det typisk også, at:

1. Mennesker afholder sig fra ved handling eller uvirksomhed at skade andre mennesker.
2. Mennesker adlyder rimelige og retfærdige instruktioner fra samfundets anerkendte autoriteter (som læger, lærere m.v.).
3. Mennesker så vidt muligt at undgår selv at pådrage sig skade.

I sin novelle "Evidence" lader Asimov en af sine karakterer gøre rede for disse sociale adfærdsregler blandt mennesker som en moralsk basis for robotlovene. I grove træk og med den forskel, at mennesker ikke bør kunne beordres til at selvdestruere eller påføre sig selv alvorlig skade, svarer Asimovs robotlove med andre ord til almindelige adfærdsregler blandt mennesker.

Robotter som redskaber?

Hvis robotlovene bliver en del af alle kunstigt intelligente systemers design, er ideen, at vi ikke længere behøver at frygte scenarier som i "The Matrix" og "The Terminator", hvor robotterne overtager verden og undertrykker menneskene i den. Som Asimov i sit essay "Robot visions" har pointeret, svarer det i virkeligheden til, hvordan de fleste redskaber er designet:

1. Et redskab skal være sikkert at bruge (knive har for eksempel skæfter, cykler til småbørn har støttehjul mv.).
2. Et redskab skal fungere effektivt, medmindre dette vil skade brugeren.
3. Et redskab skal forblive intakt under brug, medmindre dets destruktion er påkrævet for dets brug eller af sikkerhedsmæssige hensyn.

At betragte kunstigt intelligente systemer som redskaber for menneskeheden frem for ligeværdige eksistenser, har dog fået nogle til at kritisere robotlovene for at begrænse robotters fri vilje og derved gøre dem til en slags slaver.

Ifølge Singularitets-Instituttet for Kunstig Intelligens (eng: The Singularity Institute for Artificial Intelligence), der blandt andet har til formål at fremme udviklingen af sikker og gavnlig kunstig intelligens, bør man skelne mellem ”redskabsniveau kunstig intelligens” og ”ægte bevidsthed” eller ”bevidsthedsniveau kunstig intelligens”.

At bygge et intelligent, bevidst system kan ifølge instituttet ikke sammenlignes med at bygge et redskab. Redskaber kan bruges til et hvilket som helst formål, som brugeren ønsker. En bevidsthed kan derimod have selvstændige mål og kan udføre bevidste handlinger for at opnå disse mål. Samtidig mener de, det er en grundlæggende fejl at forestille sig, at kunstig intelligens vil ligne menneskelig intelligens. Mennesker har en kompleks og indviklet indre arkitektur, som er resultatet af millioner af års udvikling. Det er derfor fejlagtigt at tro, at kunstig intelligens vil fungere som menneskelig intelligens.